

# Shrinkage Estimation and Selection for Multiple Functional Regression

Heng Lian

Division of Mathematical Sciences  
School of Physical and Mathematical Sciences  
Nanyang Technological University  
Singapore, 637371  
E-mail: henglian@ntu.edu.sg

## Abstract

Functional linear regression is a useful extension of simple linear regression and has been investigated by many researchers. However, functional variable selection problems when multiple functional observations exist, which is the counterpart in the functional context of multiple linear regression, is seldom studied. Here we propose a method using group smoothly clipped absolute deviation penalty (gSCAD) which can perform regression estimation and variable selection simultaneously. We show the method can identify the true model consistently and discuss construction of pointwise confidence interval for the estimated functional coefficients. Our methodology and theory is verified by simulation studies as well as an application to spectrometrics data.

*Keywords:* Estimation consistency, functional linear regression, group SCAD, principal component analysis, selection consistency.

## 1 Introduction

In several applications, functional data appear as the basic unit of observations. Classical regression models may be inadequate for such cases because of the high correlations for the discretized data. Compared with the discrete multivariate analysis, functional analysis takes into account the smoothness of the high dimensional covariates, and often suggests new approaches to the problems that have not been discovered before. Some recent developments in functional regression include Yao et al. (2005); Cai and Hall (2006); Crambes et al. (2009); Yuan and Cai (2010+).

The literature contains an impressive range of functional analysis tools for various problems. The more traditional approach, carefully documented in the monograph Ramsay and Silverman (2005), typically starts by representing functional data by an expansion with respect to a certain basis, and subsequent inferences are carried out on the coefficients. Another line

of work by the French school, taking a nonparametric point of view, extends the traditional nonparametric techniques, most notably the kernel estimate, to the functional case (Ferraty and Vieu, 2006). Besides these two major approaches, other methods such as putting functional regression in the reproducing kernel Hilbert space framework has been developed (Preda, 2007; Lian, 2007).

In this paper, we are concerned with an extension of simple functional linear regression model to the case where multiple functional observations are made on each unit. Formally, the model we consider is

$$Y_i = a + \sum_{j=1}^p \int_0^1 \beta_j(t) X_{ij}(t) dt + \epsilon_i, 1 \leq i \leq n, \quad (1)$$

where  $X_{ij}$  are random functions,  $a$  is the intercept,  $\epsilon_i$  are random scalar errors and the functional coefficients  $\beta_j, 1 \leq j \leq p$  are the objects of interest in the model.

Because functional coefficients are more complicated objects than scalar coefficients in classical multiple linear regression, it is generally desirable to identify those significant variables in predicting the responses, even if  $p$  is small. For example, in a similar context, Zhu et al. (2010+) investigated fluorescence spectroscopy for cervical precancer diagnosis, where a Bayesian model is used to select from multiple fluorescence spectra for classification of subjects.

In a non-Bayesian context, traditional methods for variable selection in classical linear models include constructing hypothesis tests or using information criteria. More recently, regularization methods have received much attention. For standard linear regression, Lasso (Tibshirani, 1996) is probably the most popular method that uses  $L_1$  penalty to force some of the coefficients to be equal to zero. Several subsequent works (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006) have shown that Lasso is in general not consistent for model selection unless some nontrivial conditions on the covariates are satisfied. To address these shortcomings of Lasso, Fan and Li (2001) proposed the smoothly clipped absolute deviation (SCAD) penalty and Zou (2006) proposed the adaptive lasso in the fixed  $p$  case using a weighted  $L_1$  penalty with weights determined by an initial estimator, and there are many other extensions of the regularization framework for variable selection (Yuan and Lin, 2006; Wang and Leng, 2007; Huang et al., 2008).

In this article we use functional principle component analysis (PCA)-based estimation method (Cardot et al., 1999; Hall and Horowitz, 2007) combined with group SCAD (this terminology seems to first appear in Wang et al. (2007) for varying coefficients variable selection) which represents a new application of the SCAD penalty. Regularization method for variable selection in nonparametric settings has been developed in the context of smoothing spline ANOVA for nonparametric regression smoother (Lin and Zhang, 2006) and support vector machines (Zhang, 2006). For varying-coefficient models, we are aware of the work (Wang et al., 2008) where the authors used basis expansion approach for estimation combined with group SCAD penalty on coefficients, and the work (Wang and Xia, 2009) where group Lasso penalty is applied directly to smooth functions evaluated at sampled points.

The rest of the article is organized as follows. We describe the functional PCA and shrinkage estimation procedure in Section 2.1, and present estimation consistency and selection consistency results in Section 2.2. Then we discuss estimation and inference algorithms and tuning parameter selection in Section 2.3 and 2.4 respectively. In Section 3, we present some simulation experiments and illustrate the proposed method using the spectrometrics data example where the goal is to predict the percentage of fat content in the piece of meat based on spectra curves. All technical details are gathered in the Appendix.

## 2 Methodology and Theoretical Properties

### 2.1 Estimation of Multiple Functional Regression

Assume we have independent and identically distributed (i.i.d.) observations  $(X_{i1}, \dots, X_{ip}, Y_i)$ ,  $1 \leq i \leq n$ , where  $X_{ij}$  is a square integrable random function on the interval  $[0, 1]$  with mean  $\mu_j$ . The response variables  $Y_i$  are generated by model (1) with i.i.d. errors  $\epsilon_i$  of finite second moments. We also assume the errors are independent of the predictors. We will use  $(X_1, \dots, X_p, Y)$  to denote the generic random variables with distribution the same as  $(X_{i1}, \dots, X_{ip}, Y_i)$ . Let  $K_j(s, t) = \text{Cov}\{X_j(s), X_j(t)\}$ , and by Mercer's theorem we have the spectral expansion

$$K_j(s, t) = \sum_{k=1}^{\infty} \lambda_{jk} \phi_{jk}(s) \phi_{jk}(t),$$

where  $\lambda_{j1} > \lambda_{j2} > \dots > 0$  are the eigenvalues of the linear operator associated with  $K_j(s, t)$  with corresponding eigenfunctions  $\phi_{jk}$ . On the other hand, let  $\hat{K}_j(s, t) = \frac{1}{n} \sum_{i=1}^n (X_{ij}(s) - \bar{X}_j(s))(X_{ij}(t) - \bar{X}_j(t))$  where  $\bar{X}_j = \sum_i X_{ij}/n$ , we have the empirical counterpart of the above expansion,

$$\hat{K}_j(s, t) = \sum_{k=1}^{\infty} \hat{\lambda}_{jk} \hat{\phi}_{jk}(s) \hat{\phi}_{jk}(t),$$

where  $\hat{\lambda}_{j1} \geq \hat{\lambda}_{j2} \geq \dots \geq 0$ . To get rid of uncertainty of signs, we assume  $\int \hat{\phi}_j \phi_j \geq 0$ . For the empirical operator  $\hat{K}$ , at most  $n$  eigenvalues are strictly positive.

In general, different functional predictors are not independent of each other. The Karhunen-Loève expansion of the random function  $X_{ij}$  in terms of the orthonormal basis  $\phi_{jk}(t)$  is given by

$$X_{ij} - \mu_j = \sum_{k=1}^{\infty} \xi_{ijk} \phi_{jk}, \quad (2)$$

where  $\xi_{ijk}$  are principal component scores satisfying  $E\xi_{ijk} = 0$ ,  $E\xi_{ijk}^2 = \lambda_{jk}$  and  $E\xi_{ijk}\xi_{ijk'} = 0$ ,  $k \neq k'$ . Thus from (2) we have the covariance operator expansion

$$\text{Cov}\{X_{j_1}(s), X_{j_2}(t)\} = \sum_{k_1, k_2=1}^{\infty} \lambda_{k_1, k_2}^{j_1, j_2} \phi_{j_1 k_1}(s) \phi_{j_2 k_2}(t),$$

where  $\lambda_{k_1, k_2}^{j_1, j_2} = E\xi_{j_1 k_1} \xi_{j_2 k_2}$  determines the dependency structure between different predictors. Note with our notation, when  $j_1 = j_2 = j$ ,  $\lambda_{k_1, k_2}^{j, j} = 0$  if  $k_1 \neq k_2$  and  $\lambda_{k_1, k_2}^{j, j} = \lambda_{jk}$  if  $k_1 = k_2 = k$ . An illustration of how this dependency could arise is given in the next subsection.

The model (1) can be equivalently written as

$$Y_i - \mu = \sum_{j=1}^p \int \beta_j (X_{ij} - \mu_j) + \epsilon_i, \quad (3)$$

where  $\mu = E[Y|X_1, \dots, X_p] = a + \sum_j \int \beta_j \mu_j$ . After  $\beta_j$  is estimated by  $\hat{\beta}_j$ , say, the intercept  $a$  can be easily estimated by  $\hat{a} = \bar{Y} - \sum_j \int \hat{\beta}_j \bar{X}_j$ , where  $\bar{Y} = \sum_i Y_i/n$ .

Now we consider the problem of estimating  $\beta_j$ . Using the orthonormal basis  $\{\phi_{jk}\}$ , (3) can be equivalently written as

$$Y_i - \mu = \sum_{j=1}^p \sum_{k=1}^{\infty} \xi_{ijk} b_{jk} + \epsilon_i,$$

making use of the expansion  $\beta_j = \sum_k b_{jk} \phi_{jk}$ . This suggests the estimator

$$\{\hat{b}_{jk}\} = \arg \min \sum_{i=1}^n (Y_i - \bar{Y} - \sum_{j=1}^p \sum_{k=1}^K \hat{\xi}_{ijk} b_{jk})^2,$$

and then  $\hat{\beta}_j = \sum_{k=1}^K \hat{b}_{jk} \hat{\phi}_{jk}$ , where in the above displayed equation  $\hat{\xi}_{ijk} = \int (X_{ij} - \bar{X}_j) \hat{\phi}_{jk}$  is the principal component score estimated from data. Here the truncation point  $K$  is a smoothing parameter. To further select functional predictors simultaneously, we minimize the criterion function

$$J(b) = \sum_{i=1}^n (Y_i - \bar{Y} - \sum_{j=1}^p \sum_{k=1}^K \hat{\xi}_{ijk} b_{jk})^2 + n \sum_{j=1}^p p_{\lambda}(\|b_j\|), \quad (4)$$

where  $\|b_j\|$  is the  $l_2$  norm of  $b_j = (b_{j1}, \dots, b_{jK})^T$ . Among many ways to specify the penalty function  $p_{\lambda}$ , we choose the SCAD penalty function of Fan and Li (2001), which can be defined by

$$p'_{\lambda}(\theta) = \lambda \left\{ I(\theta \leq \lambda) + \frac{(a\lambda - \theta)_+}{(a-1)\lambda} I(\theta > \lambda) \right\}, \quad p_{\lambda}(0) = 0,$$

for  $a = 3.7$  and  $\theta > 0$ , where  $I(\cdot)$  is the indicator function. The choice of  $a = 3.7$  is suggested by Fan and Li (2001) and adopted in almost all publications involving SCAD penalty. Other penalty functions such as the adaptive Lasso can also be used here and will lead to similar consistency results as below.

## 2.2 Consistency Properties

Large sample properties of shrinkage estimation with SCAD penalty have been established in the literature (Fan and Li, 2001; Fan and Peng, 2004; Wang et al., 2008). We show that in our context the estimation procedure can consistently estimate the functional coefficients as well as consistently identify the true model. However, extending these theoretical results to multiple functional regression is not trivial. Note that in criterion (4) two types of approximations are involved, one is the truncation of  $\beta_j$  to approximate the functional coefficients, the other is the unknown covariate  $\xi_{ijk}$  estimated by  $\hat{\xi}_{ijk}$ . While the former approximation is typical in nonparametric problems such as Wang et al. (2008), the latter is unique to the functional regression problem. It also resembles the measurement error model in form where the covariates are not observed directly (Liang and Li, 2009; Carroll et al., 2009).

We denote the true regression coefficients by  $\beta = ((\beta^{(1)})^T, (\beta^{(2)})^T)^T$  with  $\beta^{(1)} = (\beta_1, \dots, \beta_s)^T$ ,  $s \leq p$  containing all nonvanishing components of  $\beta$  and  $\beta_{s+1} = \dots = \beta_p \equiv 0$ . Let  $\Lambda$  be the  $pK \times pK$  matrix

$$\begin{pmatrix} \Lambda^{1,1} & \dots & \Lambda^{1,p} \\ \vdots & \vdots & \vdots \\ \Lambda^{p,1} & \dots & \Lambda^{p,p} \end{pmatrix}, \quad (5)$$

where  $\Lambda^{j_1, j_2}$  is the  $K \times K$  matrix with entries  $\lambda_{k_1, k_2}^{j_1, j_2}$ ,  $1 \leq k_1, k_2 \leq K$ . In our results, the following regularity conditions are needed:

- (c1)  $X_{ij}$  has finite fourth moment:  $\int E(X_{ij}^4) < \infty$ , and  $E\epsilon_i^4 < \infty$ .
- (c2)  $\lambda_k - \lambda_{k+1} \geq C^{-1}k^{-\alpha-1}$ ,  $b_{jk} \leq Ck^{-\beta}$ ,  $\alpha > 1$ ,  $\beta > \alpha + 1/2$ .
- (c3) The smoothing parameter  $K$  satisfies  $n^{-1/(5\alpha+3)}K \rightarrow 0$ , and  $n^{-1/(3\alpha+2\beta+2)}K$  is bounded away from 0.
- (c4) The parameter  $\lambda$  satisfies  $\lambda = o(K^{-\alpha})$  and  $\sqrt{K^{3\alpha+3}/n} = o(\lambda)$ .
- (c5) The minimum eigenvalue of  $\Lambda$ , denoted by  $\rho_{\min}(\Lambda)$ , is of order  $\Omega(K^{-\alpha})$  where  $a_n = \Omega(b_n)$  means  $b_n = O(a_n)$ .

**Remark 1** Because  $\beta > \alpha + 1/2$ , there exists  $K$  satisfying (c3). Also, we have  $\sqrt{K^{3\alpha+3}/n} = o(K^{-\alpha})$  when  $K$  satisfies (c3) and thus there exists  $\lambda$  satisfying (c4).

Based on the assumptions listed above, we can establish the following result.

**Theorem 1** Assume (c1)-(c5), we have

- (a) (Estimation consistency)  $\|\hat{\beta}_j - \beta_j\| = o_p(1)$ ,  $1 \leq j \leq p$ .
- (b) (Selection consistency)  $\hat{\beta}_{s+1} = \dots = \hat{\beta}_p \equiv 0$  with probability converging to 1.

Note that the study of optimal convergence rates for multiple functional regression problem is more complicated and we do not attempt it here.

**An illustration.** Let  $p = 2$ . Suppose the eigenvalues of  $K_1$  and  $K_2$  satisfy  $\lambda_{jk} = Ck^{-\alpha}, j = 1, 2$ . If  $X_1$  and  $X_2$  are independent, then the matrix  $\Lambda$  defined in (5) is diagonal and its minimum eigenvalue is obviously of order  $\Omega(K^{-\alpha})$ . In general,  $\Lambda$  can be written as a block matrix

$$\Lambda = \begin{pmatrix} E & F \\ F^T & G \end{pmatrix},$$

where  $E$  and  $G$  are  $K \times K$  diagonal matrices containing the eigenvalues of  $K_1$  and  $K_2$  respectively. It is easy to see that the minimum eigenvalue of  $\Lambda$  is no bigger than  $CK^{-\alpha}$ , since  $\Lambda$  is similar to

$$\tilde{\Lambda} = \begin{pmatrix} E & 0 \\ 0 & G - F^T E^{-1} F \end{pmatrix}$$

(that is  $\Lambda = P^{-1} \tilde{\Lambda} P$ , for some invertible matrix  $P$ ), and obviously the eigenvalues of  $G - F^T E^{-1} F$  are dominated by those of  $G$ . In assumption (c5), we assume that the minimum eigenvalue of  $\Lambda$  is still of order  $K^{-\alpha}$  as in the independent case. This assumption thus can be thought of as a constraint on the dependence of different predictors. However, we show in the following setup this assumption is quite natural. Suppose the random functions  $X_1$  and  $X_2$  are specified by

$$X_1 = \sum_{j=1}^l a_{1j} W_j, X_2 = \sum_{j=1}^l a_{2j} W_j, \quad (6)$$

where  $W_j, 1 \leq j \leq l$  are independent mean zero random functions with Karhunen-Loève expansion give by  $W_j = \sum_k \omega_{jk} \phi_k$  (note that we assume the eigenfunctions are common to all  $W_j$ ) with  $E\omega_{jk}^2 = \kappa_{jk} > 0$ . The following proposition gives a sufficient condition under which  $\rho_{\min}(\Lambda) = \Omega(K^{-\alpha})$ .

**Proposition 1** *Suppose  $ck^{-\alpha} \leq \kappa_{jk} \leq Ck^{-\alpha}, j = 1, \dots, l$ , for some constants  $C \geq c > 0$ . If  $\{a_{1j}\}, \{a_{2j}\}$  are two fixed sequences and one is not a scalar multiple of the other (that is, for any constant  $\gamma$ , we cannot have  $a_{1j} = \gamma a_{2j}$  or  $a_{2j} = \gamma a_{1j}$  for all  $j$ ), then  $\rho_{\min}(\Lambda) = \Omega(K^{-\alpha})$ .*

### 2.3 Computation and Inferences

One can express the criterion function  $J(b)$  in vector and matrix form. Let

$$\hat{Z}_j = \begin{pmatrix} \hat{\xi}_{1j1} & \cdots & \hat{\xi}_{1jK} \\ \vdots & \vdots & \vdots \\ \hat{\xi}_{ij1} & \cdots & \hat{\xi}_{ijK} \\ \vdots & \vdots & \vdots \\ \hat{\xi}_{nj1} & \cdots & \hat{\xi}_{njK} \end{pmatrix},$$

$\hat{Z} = (\hat{Z}_1, \dots, \hat{Z}_p)$ ,  $b = (b_{11}, \dots, b_{1K}, b_{21}, \dots, b_{pK})^T$ ,  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ , the criterion (4) is written as

$$J(b) = \sum_{i=1}^n (\mathbf{Y} - \bar{Y}\mathbf{1} - Zb)^T (\mathbf{Y} - \bar{Y}\mathbf{1} - Zb) + n \sum_{j=1}^p p_\lambda(\|b_j\|), \quad (7)$$

where  $\mathbf{1}$  is the  $n$ -dimensional vector with all components ones.

We use the local quadratic approximation idea (Fan and Li, 2001) to optimize the criterion. Specifically, if  $\hat{b}^{(m)}$  is the estimate obtained in the  $m$ -th iteration, then the criterion (7) can be locally approximated by

$$\sum_{i=1}^n (\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b)^T (\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b) + \frac{n}{2} \sum_j b_j^T R(\hat{b}^{(m)}) b_j, \quad (8)$$

where  $R(\hat{b}^{(m)}) = \text{diag}\{(p'_\lambda(\|\hat{b}_1^{(m)}\|)/\|\hat{b}_1^{(m)}\|)I_K, \dots, (p'_\lambda(\|\hat{b}_p^{(m)}\|)/\|\hat{b}_p^{(m)}\|)I_K\}$  and  $I_K$  is the  $K \times K$  identity matrix. The minimizer of (8) is then given by

$$\hat{b}^{(m+1)} = (\hat{Z}^T \hat{Z} + nR(\hat{b}^{(m)}))^{-1} \hat{Z}^T (\mathbf{Y} - \bar{Y}\mathbf{1}).$$

We iterate these steps until convergence and obtain the final estimate  $\hat{b}$ . During the iterations, depending on the choice of  $\lambda$ , if some  $\|b_j\|$  is smaller than a threshold ( $10^{-5}$  in our implementation), then we set  $b_j = 0$  and ignore the corresponding predictor in future iterations.

Now we consider the construction of pointwise confidence intervals for  $\beta_j$ . Following Fan and Li (2001), the sandwich formula can be used as an estimator for the variance of the nonzero components of  $\hat{b}$ , denoted by  $\hat{b}^{(1)}$  henceforth. The estimator of asymptotic variance is given by

$$\widehat{Cov}(\hat{b}^{(1)}) = ((\hat{Z}^{(1)})^T \hat{Z}^{(1)} + nR^{(1)})^{-1} (\hat{Z}^{(1)})^T \widehat{Cov}(Y) \hat{Z}^{(1)} ((\hat{Z}^{(1)})^T \hat{Z}^{(1)} + nR^{(1)})^{-1},$$

where  $\hat{Z}^{(1)}$  denotes the selected columns of  $\hat{Z}$  corresponding to nonvanishing  $\|b_j\|$ ,  $R^{(1)}$  denotes the selected rows and columns of  $R(\hat{b})$  in a similar way, and  $\widehat{Cov}(Y)$  is the  $n \times n$  diagonal matrix with estimated squared residuals on the diagonal. The diagonal blocks of  $\widehat{Cov}(\hat{b}^{(1)})$  gives the asymptotic variance for nonvanishing  $\hat{b}_j$ .

Since  $\hat{\beta}_j(t) = \hat{b}_j^T \hat{\phi}_j(t)$ ,  $\hat{\phi}_j(t) = (\hat{\phi}_{j1}(t), \dots, \hat{\phi}_{jK}(t))^T$ . We have the natural estimator for the asymptotic variance of  $\beta_j(t)$ :

$$\widehat{Cov}(\beta_j(t)) = \hat{\phi}_j(t)^T \widehat{Cov}(\hat{b}_j) \hat{\phi}_j(t).$$

Note that here we ignored the uncertainty of  $\hat{\phi}_j$  which is also estimated from observations. However, we think this might be a reasonable first approximation because one might argue that  $\phi_{jk}$  is easier to estimate than  $\beta_j$  since  $\phi_{jk}$  needs to be estimated first and its accuracy affect the subsequent estimations. Of course this is just a heuristic argument and we later use our simulation experiments to illustrate the performance the asymptotic variance for-

mula. Estimates of the asymptotic variance can be used to construct pointwise confidence intervals for  $\beta_j(t)$  for nonzero components of the functional coefficients. Strictly speaking, the constructed intervals will only be for the truncated  $\beta_j(t)$  at cutoff  $K$  in the expansion. Thus the constructed intervals will have lower than targeted coverage rate: on the one hand the variability in  $\hat{\phi}_{jk}$  is ignored, on the other hand the interval is only for truncated functional coefficients. The bias caused will be seen from our numerical results.

## 2.4 Tuning Parameter Selection

For implementation of our method, we need to choose two smoothing parameters, the truncation point  $K$  and the regularization parameter  $\lambda$  for group SCAD penalty. In fact, for different predictors we might choose a different truncation point, which nevertheless would lead to a significant increase in computational burden, and thus we constrain the truncation point to be the same for all predictors. Besides, one can argue that the sensitivity of the estimator to the choice of  $K$  is reduced by the extra smoothing parameter  $\lambda$  in regularized estimation.

We use generalized cross-validation (GCV) to select both  $K$  and  $\lambda$ . GCV can be thought of as a short-cut for leave-one-out cross-validation, and also comes with advantageous properties (Wahba, 1990). The GCV criterion is defined by

$$GCV(K, \lambda) = \frac{1}{n} \frac{\|\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{\mathbf{Y}}\|^2}{(1 - \text{tr}(H(K, \lambda))/n)^2},$$

where  $\hat{\mathbf{Y}} = H(K, \lambda)(\mathbf{Y} - \bar{Y}\mathbf{1})$  is the fitted response values and  $H(K, \lambda) = \hat{Z}(\hat{Z}^T \hat{Z} + nR(\hat{b}))^{-1} \hat{Z}^T$  is the hat matrix.

## 3 Numerical Experiments

### 3.1 Simulation Examples

We perform a Monte Carlo experiment to investigate the finite sample performance of the estimation method, using GCV to select the two tuning parameters. Our example follow the illustration presented after Theorem 1. The simulated data is generated from model (1) with  $p = 4$  functional predictors,  $a = 0$  and the errors  $\epsilon$  distributed as  $N(0, \sigma^2)$ . For  $1 \leq j \leq 4$  independently, we take  $W_j = \sum_{k=1}^{50} \xi_{jk} \phi_k$  where  $\xi_{jk} \sim N(0, k^{-2})$ ,  $\phi_1 \equiv 1$  and  $\phi_{k+1} = \sqrt{2} \cos(k\pi t)$  for  $k \geq 1$ . Then the functional predictors are defined through the linear transformations

$$\begin{aligned} X_1 &= W_1 + \rho(W_2 + W_3), \\ X_2 &= W_2 + \rho(W_1 + W_3), \\ X_3 &= W_3 + \rho(W_1 + W_2), \\ X_4 &= W_4. \end{aligned}$$



Table 1: Simulation results for penalized multiple functional regression.

Scenario	MSE	OMSE	TP	FP	95% Cov.Prob.1	95% Cov.Prob.2
$\rho = 0.0, \sigma = 0.1$	0.73	0.64	2	0.08	0.92	0.93
$\rho = 0.2, \sigma = 0.1$	0.70	0.63	2	0.09	0.92	0.94
$\rho = 0.5, \sigma = 0.1$	0.97	0.85	2	0.08	0.93	0.94
$\rho = 0.0, \sigma = 0.3$	1.73	1.47	1.77	0.13	0.93	0.94
$\rho = 0.2, \sigma = 0.3$	2.06	1.81	1.81	0.11	0.93	0.94
$\rho = 0.5, \sigma = 0.3$	2.75	2.26	1.80	0.15	0.92	0.92

Note the scalar  $\rho$  controls the strength of dependence between different predictors, with  $\rho = 0$  resulting in independent predictors. For  $\beta_1$  and  $\beta_2$ , in terms of expansion based on  $\{\phi_k\}$ , we take  $b_1 = (-2, 1, -2, 1)^T$ ,  $b_2 = (1, -1, 0.5, -0.5)^T$  and set  $\beta_3 = \beta_4 = 0$ . We fix  $n = 100$  for all our simulations and set  $\rho = 0, 0.2$  or  $0.5$ , and  $\sigma^2 = 0.1$  or  $0.3$ . All integrations required in the generation of the data and the estimation procedure are performed using a Riemannian sum approximation with an equally spaced grid containing 500 points on  $[0, 1]$ .

The simulation results are summarized in Table 1 based on 500 runs in each scenario, where we report the mean squared errors  $\|\hat{\beta} - \beta\|^2$  using our regularized multiple functional regression model (MSE), oracle mean squared errors where the true zero coefficients are known and no shrinkage is applied (OMSE), average number of correctly identified nonvanishing coefficients (TP), average number of incorrectly identified nonvanishing coefficients (FP), empirical coverage probability of pointwise 95% confidence interval for  $\beta_1$  (95% Cov.Prob.1) and empirical coverage probability of pointwise 95% confidence interval for  $\beta_2$  (95% Cov.Prob.2). For each scenario, the empirical coverage probabilities reported are the averages over the grid  $(0.1, 0.2, \dots, 0.9)$  for  $\beta_1$  and  $\beta_2$  whenever they are estimated as nonzero coefficients.

As one can see from Table 1, the noise level clearly has a significant effect on the estimation errors as well as the average number of truly relevant predictors detected. However, the number of false positives remains at a low level even for larger noise variance. Compared to noise level, the correlation between different predictors seems to have milder effects, with estimation error increasing with correlation strength. The result also shows that confidence intervals constructed based on the sandwich formula for the asymptotic variance work surprising well, with only a small downward bias in our simulations. As an illustration, the true functions  $\beta_1$ ,  $\beta_2$  as well as their estimates when  $\rho = 0.2$  and  $\sigma = 0.1$  or  $0.3$  are plotted in Figure 1.

### 3.2 Spectrometrics Data

We illustrate our approach on the real spectrometrics dataset, which contains 215 spectra of light absorbance for meat samples as functions of the wavelengths. Because of the denseness of wavelengths at which the measurements are made, the subjects are naturally treated as

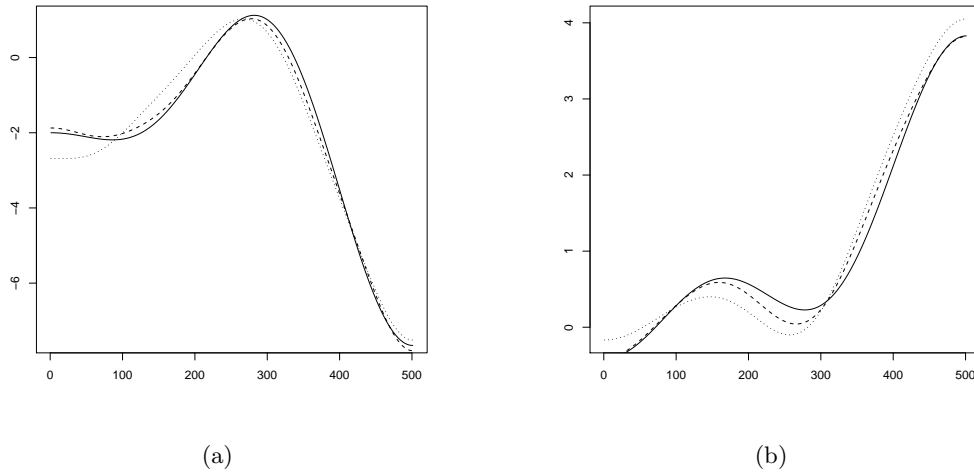
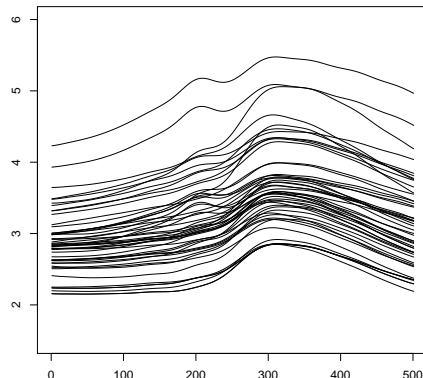


Figure 1: (a) The true coefficient  $\beta_1$  (solid line) with its estimates when  $\sigma = 0.1$  (dashed line) and  $\sigma = 0.3$  (dotted line). (b)  $\beta_2$  (solid line) with its estimates when  $\sigma = 0.1$  (dashed line) and  $\sigma = 0.3$  (dotted line). Here we set  $\rho = 0.2$ .

continuous curves. Figure 2 shows the first 50 curves in the dataset. This dataset has been previously used in functional nonparametric regression studies where the covariate is the spectra curve and the response is the percentage of fat content in the piece of meat (Ferraty and Vieu, 2002, 2006; Ferraty et al., 2007). In nonparametric kernel regression, as shown in the above mentioned works, choice of semi-metric which defined the notion of distance between curves is crucial for the performance of the estimator. Previous study suggested that for nonparametric regression function estimation, taking as the semi-metric the  $L_2$  distance between the second derivatives of the spectra gives favorable results based on its performance on hold-out validation data. A desirable feature of an estimation procedure would be to determine the appropriate order of derivative automatically.

Here we apply the multiple functional linear regression model to the spectrometrics data. We treat the original function itself as well as up to its 3rd derivatives as the predictors in our model. The idea of using different orders of derivatives of curves as covariates in the functional linear model is similar to using transformations of the original covariates in classical multiple linear regression, which makes the linear model more flexible. Compared to nonparametric functional kernel regression, the functional linear model is more easily interpretable and thus an interesting alternative. For this data, we train on the first 160 spectra and use the rest as validation. We examine the prediction accuracy of the estimated model using mean squared error on the validation data, defined as

$$MSE = \frac{1}{55} \sum_{i=161}^{215} (Y_i - \hat{Y}_i)^2.$$



(a)

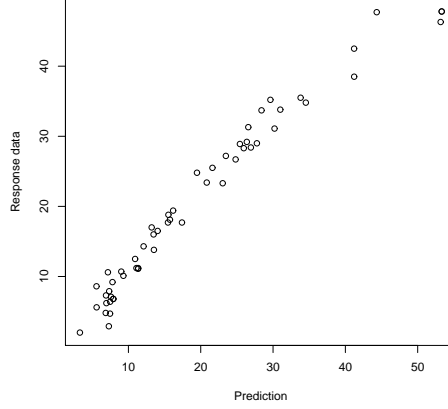
Figure 2: The spectrometric curves.

With the smoothing parameters selected by GCV, the relevant predictors are found to be the 1st and 2nd derivatives of the spectra curves, achieving an MSE of 8.31. Figure 3 clearly shows the ability of the estimated model to predict the responses. For comparison, we also computed the nonparametric kernel regression using the **funopare.kernel.cv** function provided in the **npfda** package (it uses cross-validation to select the bandwidth), which gives a smaller MSE of 5.37. However, when using functional linear modeling, unlike kernel regression, we can visually examine the features of the functional coefficients for interpretation. For example, from Figure 4, higher fat content is seen to be related to higher values around point 160 and lower values around point 215 in the first derivative, as well as lower values around point 190 in the second derivative.

## 4 Concluding Remarks

We propose in this article a regularization method for shrinkage estimation of multiple functional linear regression models. We have shown that the proposed method is consistent in estimation and variable selection. A computational algorithm based on local quadratic approximation is proposed. It is also possible to use local linear approximation (Zou and Li, 2008) and our choice made here is based on ease of implementation since closed form solution exists for each iteration. Our simulation results demonstrated the effectiveness of the method and the application to spectrometrics data provides an interesting alternative perspective to the previously used kernel regression on this data.

We would like to finish this paper by discussing some possible topics for future study. One possibility is to consider partially functional linear regression where scalar covariates are considered simultaneously. Variable selection can be applied to both the functional and non-



(a)

Figure 3: Prediction accuracy with penalized multiple functional regression on 55 validation samples.

functional part. Another direction is to consider multiple functional linear regression when the number of predictors diverges with sample size. How to extend the shrinkage estimation results to generalized functional linear model (James, 2002; Muller and Stadtmuller, 2005; Cardot and Sarda, 2005) is another interesting topic for further study.

## Appendix

The following two lemmas study some properties of the estimated principal component scores and are important for the proof of Theorem 1. Throughout the appendix, we follow the notations and assumptions in the main text.

**Lemma 1** *We have  $|\hat{\xi}_{ijk} - \xi_{ijk}| = O_p(K^{\alpha+1}/\sqrt{n})$  and  $|\sum_{i=1}^n \hat{\xi}_{ij_1k_1} \hat{\xi}_{ij_2k_2}/n - \lambda_{k_1,k_2}^{j_1,j_2}| = O_p(K^{\alpha+1}/\sqrt{n})$ .*

*Proof.* Since  $\xi_{ijk} = \int (X_{ij} - \mu_j) \phi_{jk}$  and  $\hat{\xi}_{ijk} = \int (X_{ij} - \bar{X}_j) \hat{\phi}_{jk}$ , we have  $|\hat{\xi}_{ijk} - \xi_{ijk}|^2 = O_p(\|\bar{X}_j - \mu_j\|^2 + \|\hat{\phi}_{jk} - \phi_{jk}\|^2) = O_p(K^{2\alpha+2}/n)$ , using assumption (c2) and equation (5.2) in Hall and Horowitz (2007). For the second part, we have

$$\begin{aligned}
& \frac{\sum_{i=1}^n \hat{\xi}_{ij_1k_1} \hat{\xi}_{ij_2k_2}}{n} - \lambda_{k_1,k_2}^{j_1,j_2} \\
&= \left( \frac{\sum_{i=1}^n \hat{\xi}_{ij_1k_1} \hat{\xi}_{ij_2k_2}}{n} - \frac{\sum_{i=1}^n \xi_{ij_1k_1} \xi_{ij_2k_2}}{n} \right) + \left( \frac{\sum_{i=1}^n \xi_{ij_1k_1} \xi_{ij_2k_2}}{n} - \lambda_{k_1,k_2}^{j_1,j_2} \right) \\
&=: (I) + (II).
\end{aligned}$$

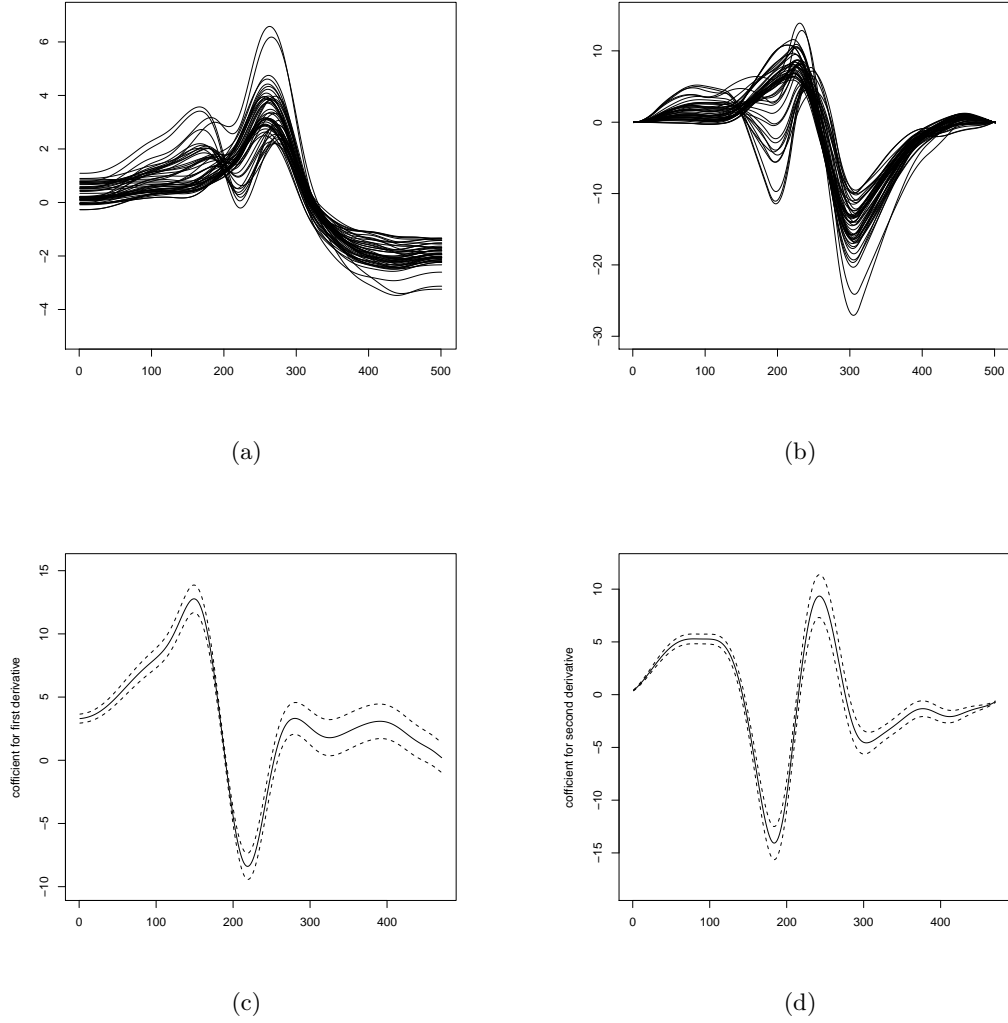


Figure 4: (a) and (b): 1st and 2nd derivative of the spectrometric data. Only 50 samples are shown in the figure. (c) and (d): Estimated functional linear coefficient corresponding to 1st and 2nd derivative curves respectively, with 95% pointwise confidence interval shown as dotted lines.

Obviously the second term is of order  $O_p(n^{-1/2})$ . The first term above is further decomposed as

$$(I) = \frac{1}{n} \sum_{i=1}^n \left[ (\hat{\xi}_{ij_1 k_1} - \xi_{ij_1 k_1}) \hat{\xi}_{ij_2 k_2} + (\hat{\xi}_{ij_2 k_2} - \xi_{ij_2 k_2}) \xi_{ij_1 k_1} \right].$$

Using  $|\hat{\xi}_{ijk}| = O_p(1)$  since  $|\hat{\xi}_{ij_1 k_1} - \xi_{ij_1 k_1}| = o_p(1)$ , we have a bound  $O_p(K^{\alpha+1}/\sqrt{n})$  for (I), and the proof is completed.  $\square$

**Lemma 2** *For any  $A$  denoting a subset of  $\{1, 2, \dots, p\}$ , let  $\hat{Z}_A$  be the columns of  $\hat{Z}$  corresponding to those predictors in  $A$ , and similarly let  $\Lambda_A$  be the submatrix of  $\Lambda$  corresponding to the predictors in  $A$ , then the minimum eigenvalue of  $\hat{Z}_A^T \hat{Z}_A / n$  is lower bounded by a constant multiple of  $K^{-\alpha}$ , i.e.  $\rho_{\min}(\hat{Z}_A^T \hat{Z}_A / n) = \Omega_p(K^{-\alpha})$ .*

*Proof.* We will use  $\|\cdot\|$  to denote also the operator norm of a matrix, and use  $\|\cdot\|_1$  for maximum row sum of a matrix. Since  $|\rho_{\min}(\hat{Z}_A^T \hat{Z}_A / n) - \rho_{\min}(\Lambda_A)| \leq \|\hat{Z}_A^T \hat{Z}_A / n - \Lambda_A\| \leq \|\hat{Z}_A^T \hat{Z}_A / n - \Lambda_A\|_1 = O_p(K^{\alpha+2}/\sqrt{n})$  using the previous lemma. This together with assumption (c5) implies the statement of the lemma.  $\square$

**Proof of Theorem 1.** In the proof we denote the minimum eigenvalue of  $\hat{Z}^T \hat{Z} / n$  by  $\rho^*$  and thus  $\rho^* = \Omega_p(K^{-\alpha})$  by Lemma 2. The true functional coefficients are denoted by  $\beta_j = \sum_k b_{jk} \phi_{jk}$ . Then

$$\begin{aligned} 0 &\geq J(\hat{b}) - J(b) \\ &= \|\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}\hat{b}\|^2 - \|\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b\|^2 + n \sum_j p_\lambda(\|\hat{b}_j\|) - n \sum_j p_\lambda(\|b_j\|) \\ &= \|\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b + \hat{Z}b - \hat{Z}\hat{b}\|^2 - \|\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b\|^2 + n \sum_j p_\lambda(\|\hat{b}_j\|) - n \sum_j p_\lambda(\|b_j\|) \\ &= 2(\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b)^T \hat{Z}(b - \hat{b}) + \|\hat{Z}(b - \hat{b})\|^2 + n \sum_j p_\lambda(\|\hat{b}_j\|) - n \sum_j p_\lambda(\|b_j\|). \end{aligned}$$

Let  $\eta = \hat{Z}(\hat{Z}^T \hat{Z})^{-1} \hat{Z}^T (\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b)$  be the projection of  $\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b$  onto the columns of  $\hat{Z}$ , Lemma 3 below shows that  $\|\eta\|^2 = O_p(r_n^2)$ , where  $r_n^2 = O_p(K^{2\alpha+3} + nK^{-\alpha-2\beta+1})$ . By assumption (c3), we get  $nK^{-\alpha-2\beta+1} = O(K^{2\alpha+3})$  and thus  $r_n^2 = O(K^{2\alpha+3})$  and the above

displayed equation can be continued as

$$\begin{aligned}
0 &\geq -O_p(r_n) \|\hat{Z}(b - \hat{b})\| + \|\hat{Z}(b - \hat{b})\|^2 + n \sum_j p_\lambda(\|\hat{b}_j\|) - n \sum_j p_\lambda(\|b_j\|) \\
&\geq -O_p(r_n^2) - \frac{1}{2} \|\hat{Z}(b - \hat{b})\|^2 + \|\hat{Z}(b - \hat{b})\|^2 + n \sum_j p_\lambda(\|\hat{b}_j\|) - n \sum_j p_\lambda(\|b_j\|) \quad (9) \\
&\geq -O_p(r_n^2) + n\rho^* \|b - \hat{b}\|^2 - n\lambda \sum_j \|\hat{b}_j - b_j\| \\
&\geq -O_p(r_n^2) + n\rho^* \|b - \hat{b}\|^2 - \frac{2n\lambda^2}{\rho^*} - \frac{n\rho^*}{2} \|\hat{b} - b\|^2,
\end{aligned}$$

where we used Cauchy-Schwartz inequality on the second line, the property  $|p_\lambda(a) - p_\lambda(b)| \leq \lambda|a - b|$  on the third line, and Cauchy-Schwartz inequality again on the last line. Thus  $\|\hat{b} - b\|^2 = O_p(\frac{r_n^2}{n\rho^*} + \frac{\lambda^2}{(\rho^*)^2}) = o_p(1)$  by assumptions (c3) and (c4).

The convergence rate for  $\|\hat{b} - b\|^2$  can be improved to  $O_p(r_n^2/n\rho^*)$ , which is useful in the proof of part (b). Since  $\|\hat{b} - b\| = o_p(1)$  and  $\lambda \rightarrow 0$ , we have  $P(p_\lambda(\|\hat{b}_j\|) = p_\lambda(\|b_j\|), 1 \leq j \leq s) \rightarrow 1$  and thus  $\sum_j p_\lambda(\|\hat{b}_j\|) - \sum_j p_\lambda(\|b_j\|) \geq 0$  with probability converging to 1. This combined with (9) gives  $\|\hat{b} - b\|^2 = O_p(r_n^2/n\rho^*)$ .

From  $\|\hat{b}_j - b_j\| = o_p(1)$ , part (a) is easily shown using the following decomposition

$$\begin{aligned}
\|\hat{\beta}_j - \beta_j\|^2 &\leq 3\|\hat{b}_j - b_j\|^2 + 3 \int \left[ \sum_{k=1}^K b_{jk}(\hat{\phi}_{jk} - \phi_{jk}) \right]^2 + 3 \sum_{k=K+1}^{\infty} b_{jk}^2 \\
&= 3\|\hat{b}_j - b_j\|^2 + 3K \sum_{k=1}^K b_{jk}^2 \|\hat{\phi}_{jk} - \phi_{jk}\|^2 + 3 \sum_{k=K+1}^{\infty} b_{jk}^2 \\
&= 3\|\hat{b}_j - b_j\|^2 + O(K \cdot \frac{K^{2\alpha+2}}{n}) + 3 \sum_{k=K+1}^{\infty} b_{jk}^2,
\end{aligned}$$

and obviously all the terms in the above display converge to zero.

Now we prove part (b) of the theorem. Let  $\hat{b}^* = (\hat{b}_1^T, \dots, \hat{b}_s^T, 0, \dots, 0)^T$ , that is,  $\hat{b}^*$  is obtained from  $\hat{b}$  by constraining the truly irrelevant components to be zero. By similar

arguments for the proof of part (a), we have

$$\begin{aligned}
0 &\geq J(\hat{b}) - J(\hat{b}^*) \\
&= 2(\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}\hat{b}^*)^T \hat{Z}(\hat{b} - \hat{b}^*) + \|\hat{Z}(\hat{b} - \hat{b}^*)\|^2 + n \sum_j p_\lambda(\|\hat{b}_j\|) - n \sum_j p_\lambda(\|\hat{b}_j^*\|) \\
&\geq -O_p(\|\eta^*\|) \|\hat{Z}(\hat{b} - \hat{b}^*)\| + n \sum_{j=s+1}^p p_\lambda(\|\hat{b}_j\|) \\
&\geq -O_p(\|\eta^*\|) \sqrt{n} \sum_{j=s+1}^p \|\hat{b}_j\| + n\lambda \sum_{j=s+1}^p \|\hat{b}_j\|, \tag{10}
\end{aligned}$$

where  $\eta^* = \hat{Z}(\hat{Z}^T \hat{Z})^{-1} \hat{Z}^T (\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}\hat{b}^*)$ . In the last line above we use the fact that  $\|\hat{b}\|_j = O_p(r_n/\sqrt{n\rho^*}) = o_p(\lambda)$  when  $j > s$  (from the proof of part (a)) and thus  $p_\lambda(\|\hat{b}_j\|) = \lambda\|\hat{b}_j\|$ .

We bound  $\|\eta^*\|$  as follows.

$$\begin{aligned}
\|\eta^*\|^2 &\leq 2\|\eta\|^2 + 2\|\hat{Z}(\hat{b}^* - b)\|^2 \\
&= O_p(r_n^2) + O_p(nr_n^2/(n\rho^*)) = O_p(r_n^2/\rho^*).
\end{aligned}$$

Since we have that  $O_p(\|\eta^*\|) = o_p(\sqrt{n}\lambda)$ , we will have a contradiction in (10) if  $\sum_{j=s+1}^p \|\hat{b}_j\| > 0$ .  $\square$

**Lemma 3** *Let  $\eta = \hat{Z}(\hat{Z}^T \hat{Z})^{-1} \hat{Z}^T (\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b)$  as in the proof of Theorem 1, then  $\|\eta\|^2 = O_p(r_n^2)$  where  $r_n^2 = K^{2\alpha+3} + nK^{-\alpha-2\beta+1}$ .*

*Proof.* Denote by  $Z$  the matrix similar in structure to  $\hat{Z}$  but contains the true principal component scores  $\xi_{ijk}$  instead of  $\hat{\xi}_{ijk}$ . We have the decomposition

$$\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b = \epsilon + (\mu - \bar{Y})\mathbf{1} + (Z - \hat{Z})b + \nu, \tag{11}$$

where  $\epsilon = (\epsilon_1, \dots, \epsilon_n)^T$  and  $\nu$  is a  $n$ -dimensional vector with  $i$ -th component given by

$$\nu_i = \sum_{j=1}^p \sum_{k=K+1}^{\infty} \xi_{ijk} b_{jk}.$$

Let  $P_{\hat{Z}} = \hat{Z}(\hat{Z}^T \hat{Z})^{-1} \hat{Z}^T$ . Now  $\eta = P_{\hat{Z}}(\mathbf{Y} - \bar{Y}\mathbf{1} - \hat{Z}b)$  is the projection of the four terms in the decomposition (11) onto columns of  $\hat{Z}$ , and we bound each term in turn below.

Since  $\|P_{\hat{Z}}\epsilon\|^2 = \epsilon^T P_{\hat{Z}} \epsilon$ , using the fact  $E[\epsilon^T P_{\hat{Z}} \epsilon | X] = \sigma^2 \text{tr}(P_{\hat{Z}}) = \sigma^2 pK$ ,  $\text{Var}(\epsilon^T P_{\hat{Z}} \epsilon | X) = 2\sigma^4 \text{tr}(P_{\hat{Z}}^2) + (E\epsilon_i^4 - 3\sigma^2) \sum_{j=1}^n (P_{\hat{Z}})_{jj}^2 \leq 2\sigma^4 \text{tr}(P_{\hat{Z}}^2) + |E\epsilon_i^4 - 3\sigma^2| \sum_{j=1}^n (P_{\hat{Z}})_{jj} = O_p(K)$  (see for example equations (3.3), (3.4) in Huang and Fan (1999)), where  $(P_{\hat{Z}})_{jj}$  are the diagonal elements of  $P_{\hat{Z}}$  which are all no larger than 1 since  $P_{\hat{Z}}$  is a projection matrix. Using the equalities  $E\epsilon^T P_{\hat{Z}} \epsilon = E[E(\epsilon^T P_{\hat{Z}} \epsilon | X)]$  and  $\text{Var}(\epsilon^T P_{\hat{Z}} \epsilon) = E[\text{Var}(\epsilon^T P_{\hat{Z}} \epsilon | X)] + \text{Var}(E[\epsilon^T P_{\hat{Z}} \epsilon | X])$ ,



we have

$$\|P_{\hat{Z}}\epsilon\|^2 = O_p(K). \quad (12)$$

Besides,

$$\|P_{\hat{Z}}(Z - \hat{Z})b\|^2 \leq \|(Z - \hat{Z})b\|^2 = O(\|(Z - \hat{Z})^T(Z - \hat{Z})\|).$$

Using Lemma 1, we get  $\|(Z - \hat{Z})^T(Z - \hat{Z})\| \leq \|(Z - \hat{Z})^T(Z - \hat{Z})\|_1 = O_p(K^{2\alpha+3})$  and thus

$$\|P_{\hat{Z}}(Z - \hat{Z})b\|^2 = O_p(K^{2\alpha+3}). \quad (13)$$

Finally,

$$\begin{aligned} \text{Var}\left(\sum_{k=K+1}^{\infty} \xi_{ijk} b_{jk}\right) &= \sum_{k=K+1}^{\infty} \lambda_{jk} b_{jk}^2 \\ &= O\left(\sum_{k=K+1}^{\infty} k^{-\alpha} k^{-2\beta}\right) \\ &= O(K^{-\alpha-2\beta+1}) \end{aligned}$$

Since the number of predictors  $p$  is fixed, we have  $\text{Var}(\nu_i) = O(K^{-\alpha-2\beta+1})$  and thus

$$\|\nu\|^2 = O_p(nK^{-\alpha-2\beta+1}). \quad (14)$$

Combining (12), (13), (14) as well as  $|\mu - \bar{Y}| = O_p(n^{-1/2})$ , we get  $\|\eta\|^2 = O_p(r_n^2)$ .  $\square$

**Proof of Proposition 1.** From equation (6), the Karhunen-Loève expansion of the predictors are given by

$$X_i(t) = \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t), \text{ with } \xi_{ik} = \sum_{j=1}^l a_{ij} \omega_{jk}, i = 1, 2.$$

Using the notation in the main text, we have that the general entries of  $\Lambda$  are given by

$$\lambda_{k_1, k_2}^{i_1, i_2} = E \xi_{i_1 k_1} \xi_{i_2 k_2} = \begin{cases} \sum_{j=1}^l a_{i_1 j} a_{i_2 j} \kappa_{jk} & k_1 = k_2 = k \\ 0 & k_1 \neq k_2. \end{cases}$$

Thus in this case, in the block matrix form,

$$\Lambda = \begin{pmatrix} E & F \\ F^T & G \end{pmatrix},$$

and the matrix  $F$  is also diagonal. Since  $\Lambda$  is similar to the matrix

$$\tilde{\Lambda} = \begin{pmatrix} E & 0 \\ 0 & G - F^T E^{-1} F \end{pmatrix},$$

the eigenvalues of  $\Lambda$  are just the diagonal elements of  $E$  and  $G - F^T E^{-1} F$ . The eigenvalues of  $E$  are  $\Omega(K^{-\alpha})$  by assumption, and the diagonal elements of  $G - F^T E^{-1} F$  are

$$\begin{aligned}
& \sum_{j=1}^l a_{2j}^2 \kappa_{jk} - \frac{(\sum_{j=1}^l a_{1j} a_{2j} \kappa_{jk})^2}{\sum_{j=1}^l a_{1j}^2 \kappa_{jk}} \\
&= \frac{\sum_{1 \leq j_1 \neq j_2 \leq l} (a_{1j_1} a_{2j_2} \sqrt{\kappa_{j_1 k} \kappa_{j_2 k}} - a_{2j_1} a_{1j_2} \sqrt{\kappa_{j_1 k} \kappa_{j_2 k}})^2}{2 \sum_{j=1}^l a_{1j}^2 \kappa_{jk}} \\
&\geq \frac{c^2 \sum_{1 \leq j_1 \neq j_2 \leq l} (a_{1j_1} a_{2j_2} - a_{2j_1} a_{1j_2})^2 k^{-2\alpha}}{2C \sum_{j=1}^l a_{1j}^2 k^{-\alpha}} \\
&= \Omega(k^{-\alpha})
\end{aligned}$$

□

## References

- Cai, T. T. and Hall, P. (2006), “Prediction in functional linear regression,” *Annals of Statistics*, 34, 2159–2179.
- Cardot, H., Ferraty, F., and Sarda, P. (1999), “Functional linear model,” *Statistics & Probability Letters*, 45, 11–22.
- Cardot, H. and Sarda, P. (2005), “Estimation in generalized linear models for functional data via penalized likelihood,” *Journal of Multivariate Analysis*, 92, 24–41.
- Carroll, R. J., Delaigle, A., and Hall, P. (2009), “Nonparametric prediction in measurement error models,” *Journal of the American Statistical Association*, 104, 993–1003.
- Crambes, C., Kneip, A., and Sarda, P. (2009), “Smoothing splines estimators for functional linear regression,” *Annals of Statistics*, 37, 35–72.
- Fan, J. Q. and Li, R. Z. (2001), “Variable selection via nonconcave penalized likelihood and its oracle properties,” *Journal of the American Statistical Association*, 96, 1348–1360.
- Fan, J. Q. and Peng, H. (2004), “Nonconcave penalized likelihood with a diverging number of parameters,” *Annals of Statistics*, 32, 928–961.
- Ferraty, F., Mas, A., and Vieu, P. (2007), “Nonparametric regression on functional data: Inference and practical aspects,” *Australian & New Zealand Journal of Statistics*, 49, 267–286.
- Ferraty, F. and Vieu, P. (2002), “The functional nonparametric model and application to spectrometric data,” *Computational Statistics*, 17, 545–564.

- (2006), *Nonparametric functional data analysis: theory and practice*, Springer series in statistics, New York, NY: Springer.
- Hall, P. and Horowitz, J. L. (2007), “Methodology and convergence rates for functional linear regression,” *Annals of Statistics*, 35, 70–91.
- Huang, J., Horowitz, J. L., and Ma, S. G. (2008), “Asymptotic properties of bridge estimators in sparse high-dimensional regression models,” *Annals of Statistics*, 36, 587–613.
- Huang, L. S. and Fan, J. Q. (1999), “Nonparametric estimation of quadratic regression functionals,” *Bernoulli*, 5, 927–949.
- James, G. M. (2002), “Generalized linear models with functional predictors,” *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 64, 411–432.
- Lian, H. (2007), “Nonlinear functional models for functional responses in reproducing kernel Hilbert spaces,” *Canadian Journal of Statistics-Revue Canadienne De Statistique*, 35, 597–606.
- Liang, H. and Li, R. Z. (2009), “Variable selection for partially linear models with measurement errors,” *Journal of the American Statistical Association*, 104, 234–248.
- Lin, Y. and Zhang, H. H. (2006), “Component selection and smoothing in multivariate nonparametric regression,” *Annals of Statistics*, 34, 2272–2297.
- Meinshausen, N. and Bühlmann, P. (2006), “High-dimensional graphs and variable selection with the Lasso,” *Annals of Statistics*, 34, 1436–1462.
- Müller, H. G. and Stadtmüller, U. (2005), “Generalized functional linear models,” *Annals of Statistics*, 33, 774–805.
- Preda, C. (2007), “Regression models for functional data by reproducing kernel Hilbert spaces methods,” *Journal of Statistical Planning and Inference*, 137, 829–840.
- Ramsay, J. O. and Silverman, B. W. (2005), *Functional data analysis*, Springer series in statistics, New York: Springer, 2nd ed.
- Tibshirani, R. (1996), “Regression shrinkage and selection via the Lasso,” *Journal of the Royal Statistical Society Series B-Methodological*, 58, 267–288.
- Wahba, G. (1990), *Spline models for observational data*, Philadelphia, PA: Society for Industrial and Applied Mathematics.
- Wang, H. S. and Leng, C. L. (2007), “Unified LASSO estimation by least squares approximation,” *Journal of the American Statistical Association*, 102, 1039–1048.
- Wang, H. S. and Xia, Y. C. (2009), “Shrinkage estimation of the varying coefficient model,” *Journal of the American Statistical Association*, 104, 747–757.

- Wang, L. F., Chen, G., and Li, H. Z. (2007), “Group SCAD regression analysis for microarray time course gene expression data,” *Bioinformatics*, 23, 1486–1494.
- Wang, L. F., Li, H. Z., and Huang, J. H. Z. (2008), “Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements,” *Journal of the American Statistical Association*, 103, 1556–1569.
- Yao, F., Muller, H. G., and Wang, J. L. (2005), “Functional linear regression analysis for longitudinal data,” *Annals of Statistics*, 33, 2873–2903.
- Yuan, M. and Cai, T. T. (2010+), “A reproducing kernel Hilbert space approach to functional linear regression,” *Annals of Statistics*, to appear.
- Yuan, M. and Lin, Y. (2006), “Model selection and estimation in regression with grouped variables,” *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 68, 49–67.
- Zhang, H. H. (2006), “Variable selection for support vector machines via smoothing spline ANOVA,” *Statistica Sinica*, 16, 659–674.
- Zhao, P. and Yu, B. (2006), “On model selection consistency of Lasso,” *Journal of Machine Learning Research*, 7, 2541–2563.
- Zhu, H., Vannucci, M., and Cox, D. (2010+), “A Bayesian Hierarchical Model for Classification with Selection of Functional Predictors,” *Biometrics*, to appear.
- Zou, H. (2006), “The adaptive lasso and its oracle properties,” *Journal of the American Statistical Association*, 101, 1418–1429.
- Zou, H. and Li, R. Z. (2008), “One-step sparse estimates in nonconcave penalized likelihood models,” *Annals of Statistics*, 36, 1509–1533.